Neural Networks – The New Moore's Law

Chris Rowen, PhD, FIEEE
CEO
Cognite Ventures

December 2016

Outline

- Moore’s Law Revisited: Efficiency Drives Productivity
- Embedded Neural Network Product Segments
- Efficiency and Productivity in Embedded Neural Networks
- Breadth of Neural Networks:
  - Vision and the Pixel Explosion
  - Speech
  - Natural Language
- Efficiency, Productivity and Neural Networks
- The Embedded Systems Innovation Space
- The Evolution of Electronic Design and Cognitive Computing

Moore's Law Revisited:

Efficiency Drives Productivity

Moore's "Law": number of transistors in a [economically viable] dense integrated circuit doubles approximately every two years

Dennard Scaling	Impact of scaling L by α (<1)
Density: Transistors and gates per unit area	$\frac{1}{\alpha^2}$
Speed: Gate delay	α
Power: Energy per switch	α^3

Dennard benefits hit by limits of voltage scaling

- Compound effect of cost and performance scaling has revolutionized electronics
- Calculations/$ improved ~10^{10} in 50 years
- "Excess efficiency" largely responsible for processor and software revolution

Cognite Ventures

How is Deep Learning like Moore's Law?

Embedded Neural Network Product Segments

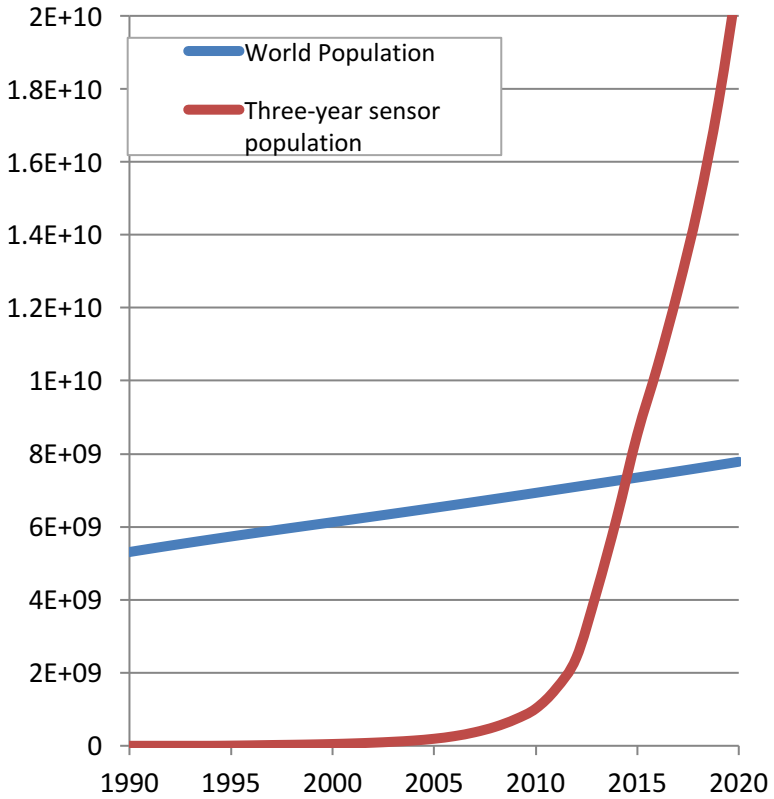
	Autonomous Vehicles and Robotics	Monitoring and Surveillance	Human-Machine Interface	Personal Device Enhancement
Vision	Multi-sensor: image, depth, speed Environmental assessment Full surround views		Attention monitoring Command interface Multi-mode ASR	Social photography Augmented Reality
Audio	Ultrasonic sensing	Acoustic surveillance Health and performance monitoring	Mood analysis Command interface	ASR social media Hands-free UI Audio geolocation
Natural Language		Access control Sentiment analysis	Mood analysis Command interface	Real-time translation Local bots Enhanced search

The Pixel Explosion

- Computing and communication driven by new data in/out
- CMOS sensors trigger imaging explosion
- 99% of of captured raw data is pixels (dwarfs sounds and motion)

10^{10} sensors x 10^{8} pixels/sec = 10^{18} raw pixels/sec

- Rapid growth of vision-based products and services
- Starting 2015: more image sensors than people
- New Age: Making sense of pixels **requires** computer cognition

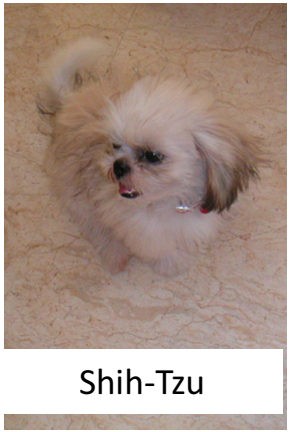


Cognite Ventures

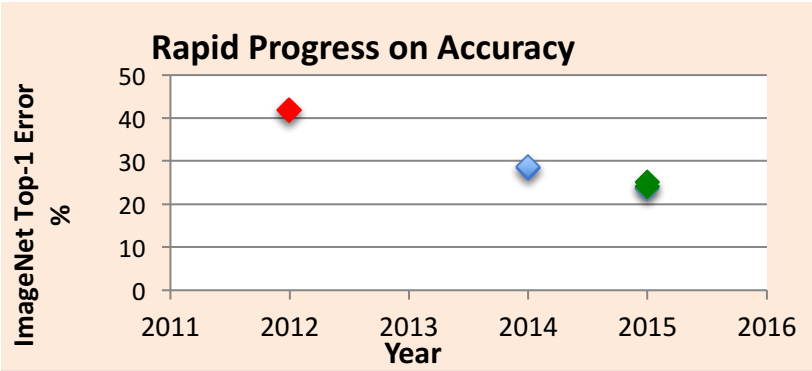
Vision

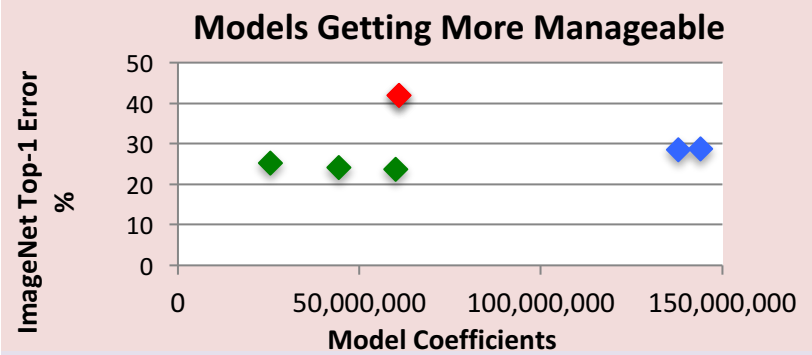
- Computer vision is big, obvious NN domain
- Many related tasks: classification, localization, segmentation, object recognition, captioning…
- Huge computation in embedded inference
- Vision is fundamentally hard!!
- Example: ImageNet Classification:
  - 1000 categories
  - 120 species of dogs

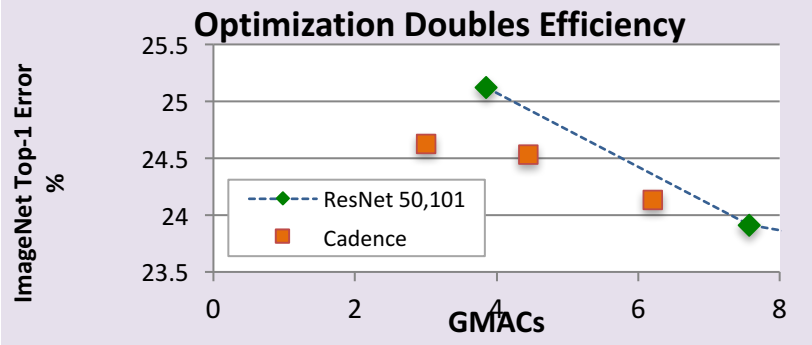

Tibetan mastiff


Shih-Tzu


Norwegian elkhound

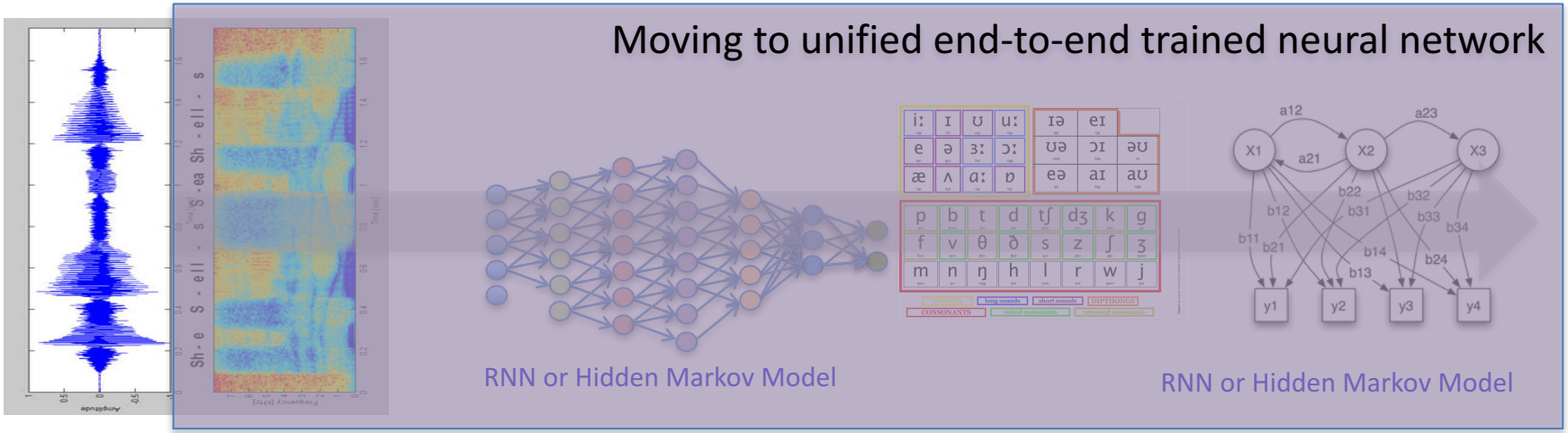






Speech: From sounds to words

- Automated speech recognition (ASR) pipeline:



"She sells seashells by the seashore"

Waveform to spectral samples

Spectral samples to phonemes

Phonemes to words

Year	Word Error Rate on "Switchboard" ASR
1995	40%
2001	20%
2015	6.3%

Tuske et al, "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR"

Natural Language

- Real time natural language interpretation crucial for rich human-machine interfaces
- But how do we automatically find word meanings??
- Powerful approach:
  - Automatically learn high-dimension vector embedding (N =300-600) from the word usage in large data-sets
  - Words whose vectors are close have strong relationship
  - Different dimensions reflect different kinds of word relationships:

    good: better

    good: fine

    good: bad

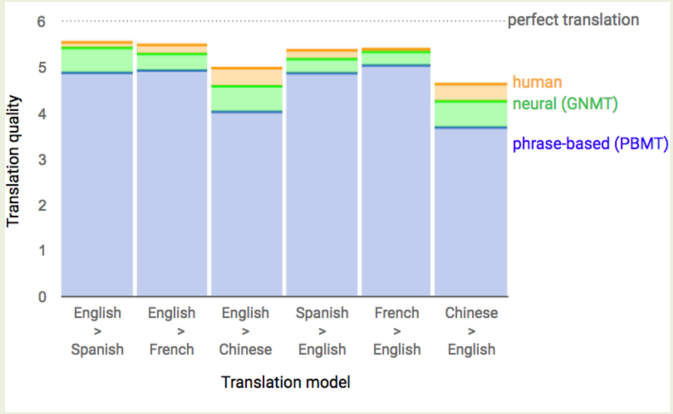
    good: product
  - Cool application: vector arithmetic reveals complex relationships:

V(“king”) – V(“man”) + V(“woman”) ≈ V(“queen”)

See: Mikolov et al, “Efficient Estimation of Word Representations in Vector Space”

Translation goes NN:

- Full sentence-at-a-time translation
- “With [Google Neural Machine Translation], Google Translate is improving more in a single leap than we’ve seen in the last ten years combined” - Barak Turovsky

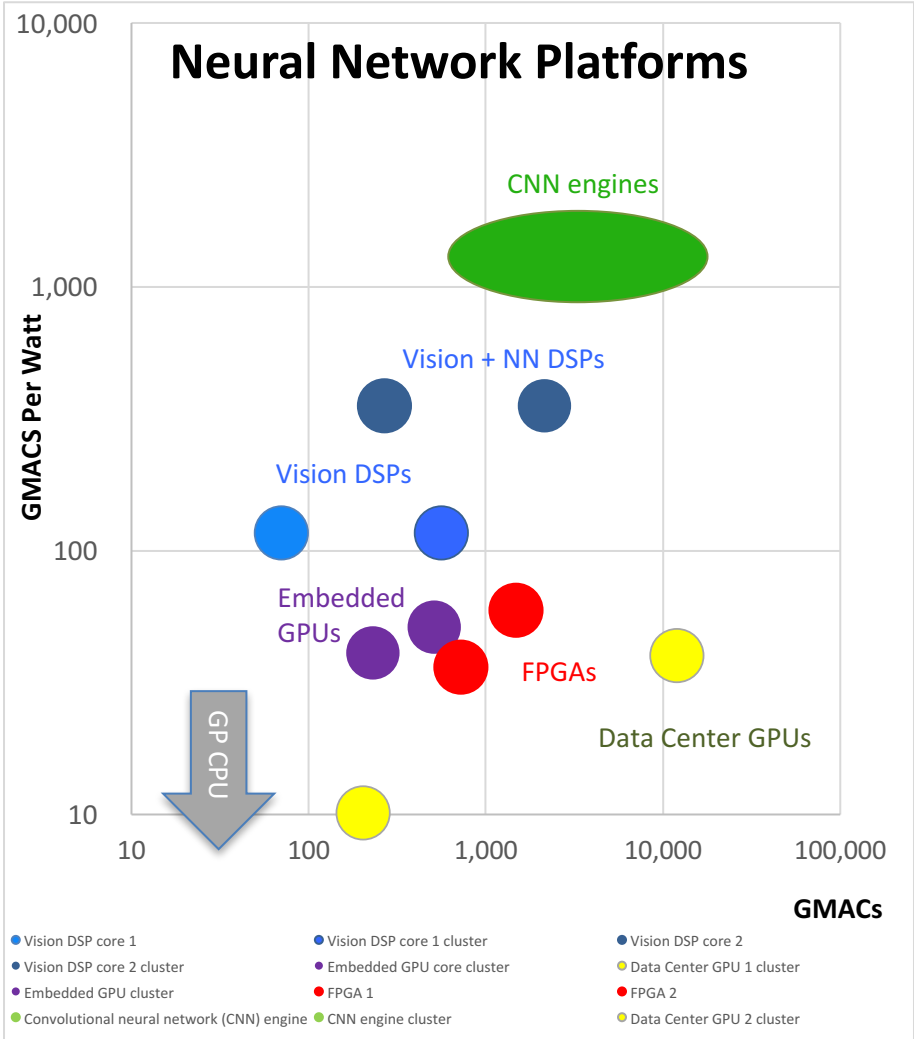


Source: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

Efficiency and Neural Networks

Efficiency:

- Conventional wisdom - deep neural networks much less efficient than hand-tuned feature recognition methods (but more effective)
- Convolutional neural networks allow
  - High parallelism
  - Low bit resolution
  - Structured, specialized architectures
  - Manageable memory bandwidth
- ~1000x energy improvement over GP CPU may compensate for efficiency gap



Cognite Ventures

Productivity and Neural Networks

"machine learning technology is on the cusp of eating software" -Alex Woodie

Productivity:

- Training neural networks often much easier than coding of application-specific features
- Neural network methods routinely perform better than best "manual" methods
- Massive educational shift underway to make machine learning a basic CS skill

MOOCs: Andrew Ng's Stanford Machine Learning course: 100,000 students registered.

- Hype still exceeds reality

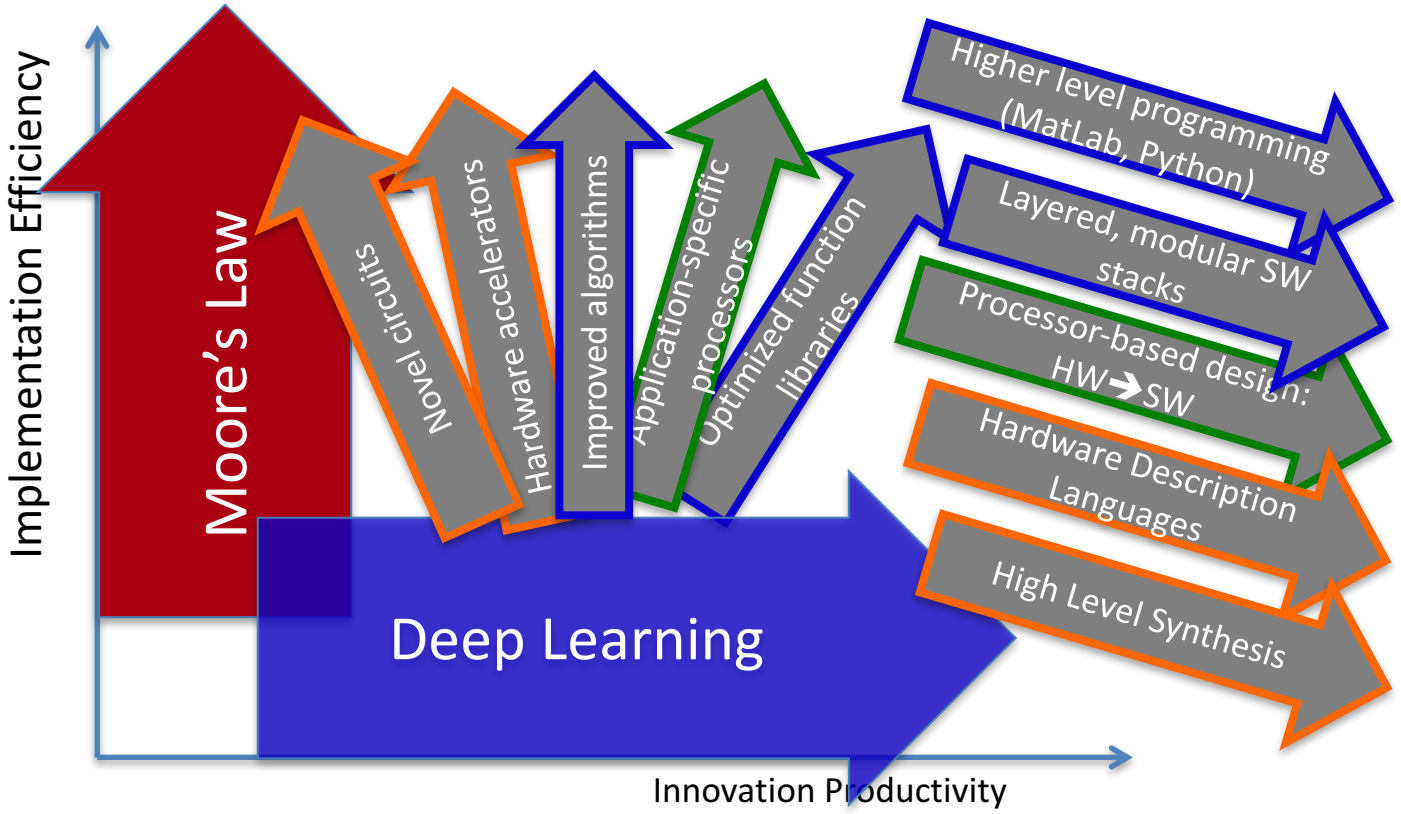
- Productivity Impact at Three Levels:

1. Programming: Trained NN replace hand-design of feature detectors
2. Applications: Rich machine learning frameworks make opens sophisticated NN to non-experts
3. Business: Deep Learning methods are enough better and faster to influence the whole tech sector.

We are in the first phase, when very rapid productivity gains become possible in parts of the economy that are simply too small to affect the overall numbers… [T]he tech sector should experience greater productivity gains over a greater range of businesses, potentially nudging measured productivity upward.

-- The Economist, July 23,2014

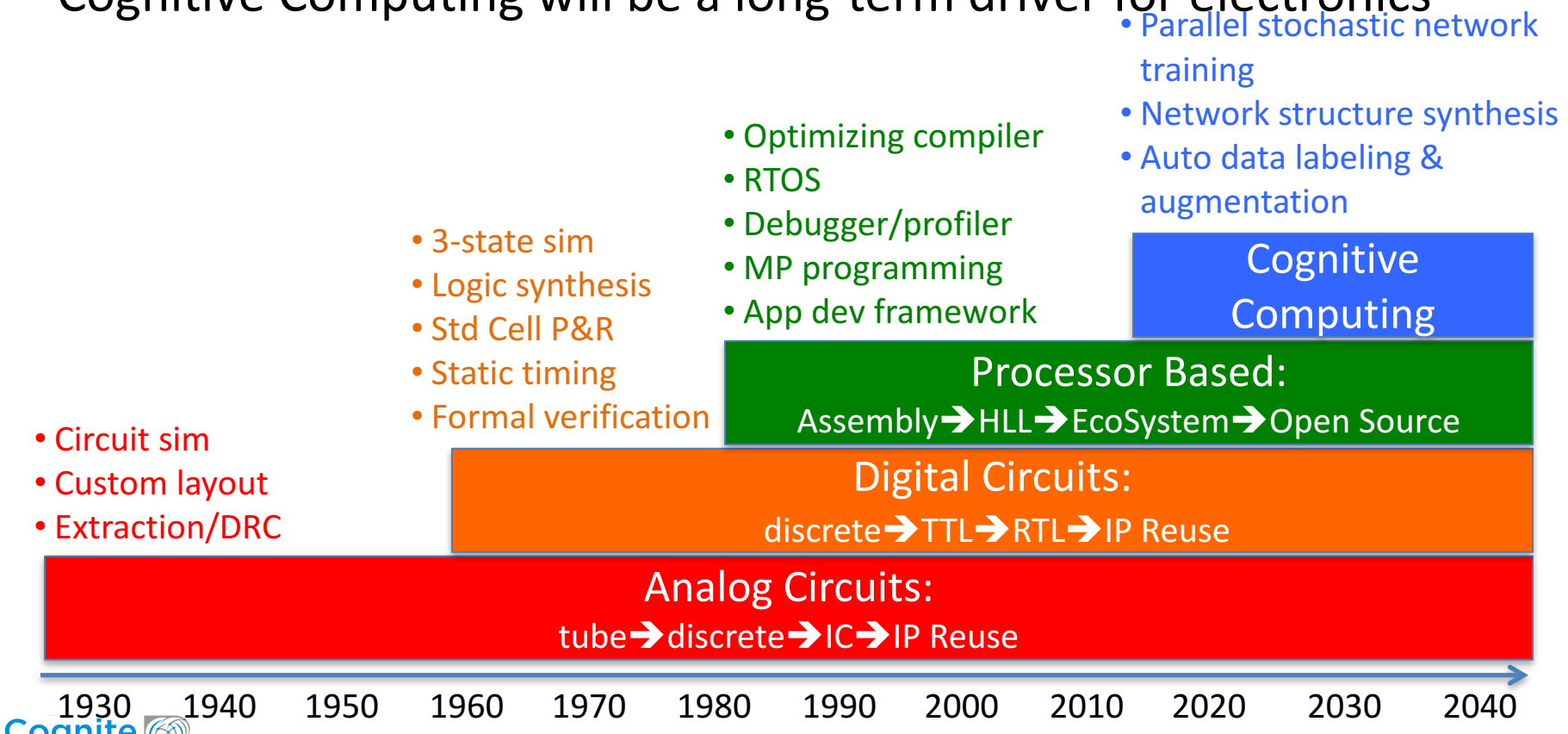
The Embedded Systems Innovation Space



Cognite Ventures

The Evolution of Electronic Design

Cognitive Computing will be a long-term driver for electronics



Cognite Ventures

neural network technology and applications